Political Geography 113 (2024) 103134

Contents lists available at ScienceDirect

Political Geography

journal homepage: www.elsevier.com/locate/polgeo

Full Length Article

A world model: On the political logics of generative AI

Louise Amoore [a,*], Alexander Campolo [a], Benjamin Jacobsen [b], Ludovico Rella [a]

[a] Department of Geography, Durham University, UK
[b] Department of Sociology, York University, UK

A B S T R A C T

The computational logics of large language models (LLMs) or generative AI – from the early models of CLIP and BERT to the explosion of text and image generation via ChatGPT and DALL-E – are increasingly penetrating the social and political world. Not merely in the direct sense that generative AI models are being deployed to govern difficult problems, whether decisions on the battlefield or responses to pandemic, but also because generative AI is shaping and delimiting the political parameters of what can be known and actioned in the world. Contra the promise of a generalizable "world model" in computer science, the article addresses how and why generative AI gives rise to a *model of the world*, and with it a set of political logics and governing rationalities that have profound and enduring effects on how we live today. The article traces the genealogies of generative AI models, how they have come into being, and why some concepts and techniques that animate these models become durable forms of knowledge that actively shape the world, even long after a specific material commercial GPT model has moved on to a new iteration. Though generative AI retains significant traces of former scientific and computational regimes – in statistical practices, probabilistic knowledge, and so on – it is also dislocating epistemological arrangements and opening them to novel ways of perceiving, characterising, classifying, and knowing the world. Four defining aspects of the political logic of generative AI are elaborated: i) *generativity* as something more than the capacity to generate image or text outputs, so that a generative logic acts upon the world understood as estimates of "underlying distributions" in data; ii) *latency* as a political logic of compression in which (by contrast with claims to reduction or distortion) the thing that is hidden, unknown or latent becomes surfaced and amenable to being governed; iii) broken and parallelized *sequences* as the ordering device of the political logic of generative AI, where attention frameworks radically change the possibilities for governing non-linear problems; iv) *pre-training and fine-tuning* as a computational logic of generative AI that simultaneously shapes a "zero shot politics" oriented towards unencountered data and new tasks. Across each of the four aspects, the article maps the emerging contemporary political logic of generative AI.

> With a single, configurable world model engine, rather than a separate model for every situation, knowledge about how the world works may be shared across tasks. (LeCun, 2022: 5).

> Knowledge takes up residence in a new space […] What event, what law do they obey, these mutations that suddenly decide that things are no longer perceived, described, expressed, characterised, classified, and known in the same way? (Foucault, 2003: 235–6).

In a 2022 essay, the Turing laureate computer scientist, Yann LeCun, set out a "path towards intelligent machines" in which he draws a familiar distinction between how humans and machines respond to what he calls an "unencountered situation" (2022: 1). Humans and animals, LeCun proposed, "know how to act in many situations they have never encountered". For example, they do not need to encounter all possible situations where something may be hot, or where an object might fall, in order to know and to act in advance of other future potential hot or falling objects. In machine learning, by contrast, "even the rarest combination of situations" must be "encountered frequently in training" – every potentially hot or falling object has to be included in a training dataset – if the system is to "avoid making dangerous mistakes when facing an unknown situation" (LeCun, 2022, p. 3). For LeCun, as for many other commentators, the pathway of future AI is defined by a new paradigm of knowledge under conditions of uncertainty, one where a machine learning model must be able to *discover* something in a data distribution, something essentially *generalizable* to new and unseen situations.

This orientation of AI towards general discovery of "how the world works" and adaptation to new domains and tasks is captured by what LeCun calls a "world model", which would supply "an internal model of how the world works" so that AI becomes "configurable" to each new situation it encounters (2022: 2–3). The powerful claim that a flexible, reconfigurable world model could deal with all potential future unencountered situations defines much of the politics of contemporary generative AI. It is a claim that promises a general resolution of difficult problems across *technical computational* and *political* paradigms: an AI model that draws upon a structure of "how the world works" in order to respond to an input it had *never encountered in training*; and a political model that is always capable of action in the face of the *unencountered*

* Corresponding author.
E-mail address: louise.amoore@durham.ac.uk (L. Amoore).

https://doi.org/10.1016/j.polgeo.2024.103134
Received 6 December 2023; Received in revised form 14 May 2024; Accepted 14 May 2024
Available online 24 May 2024

situation. Though the concept of a world model is present in the ambitions of AI designers for a better and more adaptive "fit" to the world, it is also present in the critical concerns of the humanities and social sciences, where it is said, for example, that "GPT3 does not have a model of the world" whereas "every human grows up with a model of the world" (Hayles, 2023, p. 258). In these formulations – spanning the AI proponents' desires and the critics' disquiet – a model of the world defines something that AI lacks, whether due to its inefficiencies, its absence of embodiment, or its need to incorporate and learn from prior experiences.

In this essay we begin from a different set of concerns centring precisely on how and why the technical architectures of generative AI appear to give rise to a set of distinctive political logics. We argue that large language models (LLMs) and generative AI – from the earlier models of CLIP and BERT to the explosion via ChatGPT and DALL-E – are always already instantiating a *model of the world*, and with it a set of political logics and governing rationalities that have profound and enduring effects on how we live today.[1] Contra LeCun and some of his critics, the idea of a world model does not dwell *outside* of the technical architectures of generative AI, defining its pathway, its limits, or its "fit" to the world. Rather, it is *integral* and immanent to how generative models come to make specific worlds so that a "fit" can always be approximated for any problem. Here the world model is a form of "worlding" in the sense of making present specific social and political orders (Barry, 2001; Law, 1993; Morgan, 2012). It is our wager that beginning *within* the computational and political genealogies of machine learning extends the scope for critical interventions.[2] These computational architectures of generative AI models are increasingly penetrating political architectures, not only in the direct sense that they are being deployed to respond to uncertain events – from decisions on the battlefield to pandemics – but also because they shape and delimit the ethico-political boundaries of what can be known and done in the world.

To make these boundaries intelligible we trace the genealogies of generative AI models, how they have come into being, and, crucially, why some concepts and techniques that animate these models, such as "attention", "latency", or "distribution" (among many others), become durable forms of knowledge that actively shape the world, even long after a specific material commercial GPT model has moved on to a new iteration. To be clear, there are multiple possible genealogies one could trace for generative AI, and it is in the nature of genealogical method that one could never claim to trace a single "correct" historical pathway. However, our purpose is to explain how and why the technical and political logics of generative AI have crystallized in a particular way, and what political effects and actions they make possible in the world. As Michel Foucault describes the "mobility of epistemological arrangements", past forms of knowledge become "dislocated" and opened to "mutations that suddenly decide that things are no longer perceived, described, expressed, characterized, classified, and known in the same way" (2003: 235–6). Though contemporary forms of generative AI retain significant traces of former scientific and computational regimes – in statistical practices, conditional probabilistic knowledge, and so on – they are dislocating epistemological arrangements and opening novel ways of perceiving, characterising, classifying, and knowing the world. In sum, we consider generative AI to embody a series of mutations in computational ways of knowing, and these mutations are reconfiguring political models of the world. It is our aim to address these epistemic transformations by examining the emergence of existing AI systems for the concepts, assumptions, and logics they distil and express.

In the sections that follow we elaborate four aspects of the political logic of generative AI. They are aspects in the sense that a built architecture has multiple planes or aspects that nonetheless come together to give an overall form. Though these aspects are not intended to be exhaustive, they give shape to the governing rationality of generative AI and lend conceptual substance to the actually existing technical systems through which a world model is built. First, we discuss *generativity* as more than simply the capacity to generate or produce image and text outputs, mapping a way of acting upon the world understood as estimates of "underlying probability distributions". Second, we address *latency* as a logic in which (by contrast with claims to reduction or distortion) the thing that is hidden, unknown or latent becomes surfaced and amenable to being governed. Third, we attend to *sequences* as the form of order (and the ordering device) of the political logic of generative AI. The breaking up and parallelization of sequences in generative AI is radically changing the possibilities for governing non-linear problems or paying attention to "out of sequence" events. Finally, *pre-training and fine-tuning* define a method by which these models gain a grip on new and unencountered domains, problems or tasks. This computational logic of generative AI coalesces to shape what we call a "zero shot politics". The technical logic of "zero shot" learning marks a broad transformation from broadly supervised machine learning to a world of experimentation, generalization, and the capacity to act in all unencountered situations.

1. Generativity: the political logic of underlying distributions

In 2018, a group of researchers at *OpenAI* published "Improving Language Understanding by Generative Pre-Training", introducing the now well-known series of GPT models (Radford et al., 2018). The "G" in the acronym referred to a new class of *generative* language models, which had produced promising results on natural language processing (NLP) tasks. "Our goal", they write, "is to learn *a universal representation* that transfers with *little adaptation to a wide range of tasks*" (Radford et al., 2018, p. 2, *emphasis added*). Although this relatively technical article pre-dates the breath-taking hype that subsequently enveloped this series of models, it presented to the world an embryonic but hugely powerful idea: that with sufficient data and the right computational architecture it was possible to learn a universal representation that transfers to new or unencountered tasks with little adaptation.

The political and ethical stakes present even in this early moment are not limited to the political consequences of using generative models, nor even the more nebulous vernacular understanding of "generative AI" in the sense of *generating* things, or producing images and texts as *outputs*. Of course, this sense is important, and certainly the outputs have become the primary locus of public interest and ethical concern (e.g. too "humanlike" or insufficiently "humanlike" images; too plausible or insufficiently plausible texts; the "hallucination" of untrue outputs). Here we pose a different set of ethical and political questions. Rather than mitigating harmful outputs through appeal to a notion of existing human values, we are concerned precisely with how the models' technical properties make certain forms of value and governing possible. What are their distinctive ways of estimating distributions or making predictions? How do they interpolate between data elements to form populations?

If generativity exceeds the capability of producing *outputs,* we argue that this derives from a different notion of generativity at play: that there is something yielded in the estimation of an *underlying joint probability distribution* that exceeds the sum of the parts or data elements. What do we mean by a political logic of underlying distributions? Generative models attempt to learn the underlying probability distribution of the

[1] Our use of "logic" does not imply a coherence between technical and political concepts, nor a causal relation between them. Following Mol, there is not "a shared ontology" and yet a logic emerges that is "held together" by the "resonances between" the technical and political worlds (Mol, 2003: 115). For a technology such as generative AI to have a political logic, it contributes to a broader "governing rationality" or "a form of activity aiming to shape, guide or affect the conduct of some person or persons" (Gordon, 1991: 2).

[2] Among the antecedents to a conceptual-genealogical approach are those who address how machine learning makes possible new ways of knowing and acting (McQuillan, 2018), how algorithms exist as culture or patterns of meaning (Seaver, 2018), and how models become arrangements of propositions in the world (Amoore, 2020).

data they are trained on. This means that, in place of learning the relationship between a set of specific data points and labels, generative models estimate the distribution from which the training data was drawn and sample from this estimate in order to generate new data points.

To illustrate the logic of estimating an underlying distribution it is helpful to think of a contrasting logic. Much recent progress in machine learning has been due to the construction of large, high-quality *labelled* datasets, with ImageNet serving as a paradigmatic case for image recognition deep learning (Denton et al., 2021). In discriminative modelling – to which generative modelling was conceptually opposed – the learning problem is modelling relationships between images and labels, inputs and responses: given some subsample of data, can the model learn a mapping that will allow it to accurately classify images outside of this training sample? Discriminative modelling refers to this direct mapping of a conditional distribution of inputs to labels (Ng & Jordan, 2001; Rubinstein & Hastie, 1997; Vapnik, 1998).

But what should be done in cases where such labels are unavailable or impractical to apply at scale? This was the case for the NLP researchers in 2018, who were trying to leverage huge corpora of unlabelled linguistic data. And it is increasingly the argument for the use of new models in, for example, the identification of tendencies in the vast unlabelled data of patients' medical records or welfare recipients' administrative records. Instead of modelling the conditional probability of a label given an input, a generative model estimates an underlying joint probability distribution from which the training inputs are conceived as having been sampled. In language modelling, for example, this estimate of the underlying distribution can be used to predict the next most likely word in a sequence. This is the more specific probabilistic sense in which these models generate new or emergent instances.

In the dual context of a political desire to leverage unlabelled data on populations *and* a computational drive for efficient scaling and novel inference, the *generative* model offers a distinctive new logic of distribution. What is specific to the contemporary logic of generativity is that it is geared to generate things in excess of the individual data elements on which it was trained. As Manual DeLanda observes in his philosophy of emergence and generativity, "the patterns have properties, tendencies, and capacities that are not present in the individual" interacting elements (2011: 23). We are interested in pushing this sense of distribution beyond its technical definition and into the realm of governing the kinds of emergent tendencies and capacities DeLanda describes. Generative techniques have their own political logic of distribution, distinct from the Aristotelian notions of distributive justice – who is entitled to what, or what a fair distribution would be – which preoccupied political theorists in the late twentieth century (Rawls, 1971; Walzer, 1983), and different from the "distribution of the sensible" that decides what or who can be perceived (Rancière, 2010, p. 36). The politics of distribution in generative AI stem from a different tradition, that of probability. This sense does have precedent in political thought, notably in the ways that statistical populations became the objects of government in the modern period (Foucault, 2007, pp. 108–9).

The emergence of a political logic of underlying distributions points toward a different set of epistemological and ethical stakes than have thus far emerged in the critical literature on AI and machine learning. There has been much compelling critical work on the types of classification that are in some ways characteristic of discriminative modelling: the pathologies of misclassification, the ways that proxy labels reproduce discriminatory patterns: should an individual be deemed creditworthy or uncreditworthy, likely to recidivate or not? What changes when we think less in the pragmatic terms of direct data-label relationships and more in the metaphysical terms implied by generative modelling: of approximating some type of underlying joint distribution behind the phenomenal world of appearances? What happens when the constructedness of labels gives way to structures said to be immanent in the data itself? What does the idea of inferring from an underlying distribution mean in terms of a political or governing logic? How might the technical propositions of modelling underlying distributions or estimating densities begin to structure actual political decisions?

Consider, for example, US software company Palantir's Artificial Intelligence Platform (AIP) for Defence and Military, claiming to "bring together the latest in large language models (LLMs) and cutting-edge AI to activate data and models for the most highly sensitive environments" (Palantir, 2022a). A video demo of the AIP begins with the scenario of a military operator "responsible for monitoring activity within Eastern Europe". An apparently discriminative model first is used to classify and identify an alert – "anomalous military activity detected" – drawing on labelled classifications such as the ImageNet-trained "military vehicle detector" algorithm. However, when the operator queries the AIP in natural language it is claimed that "the LLM is traversing a data foundation of real-time information integrated from across public and classified sources". The generative LLM has been trained to model complex syntactical relationships on a huge range of linguistic corpora, and is fine-tuned on "military doctrine, logistics, and battle dynamics" as well as public sources such as weather forecasting and geospatial data.

Palantir's battlefield AI is thus attempting to model the structure of an underlying joint distribution of multiple data sources, inferring from the contours of this distribution a series of strategic "courses of action" with associated probabilities. These are not determined in advance by rule, label or axiom but are generated probabilistically by the LLM as it traverses the model's "data foundation". When the human operator prompts the model to "generate three courses of action (COAs) to target this enemy equipment", the LLM outputs three strategic options for human commanders to review: "COA 1: target with air asset; COA 2: target with long range artillery; COA 3: target with a tactical team" (Palantir, 2022a).[3] These courses of action, with their profoundly political and even lethally violent consequences, are not deterministically inferred from the input. Via a "sampling" of the underlying distribution, the model is generating a genuinely novel output, albeit one that is plausible given the underlying distribution modelled by the LLM. The output takes the form not of a contingent estimation of probability, but rather in the terms set by the language modelling objective, a *desired behaviour*. Crucially in the military contexts where such generative models are fine-tuned, these outputs are trained so as to be *actionable* – three strategic "courses of action". This condensation of probabilities into three outputs also radically forecloses the potential for alternative decisions outside of empirical training data and therefore the modelled underlying distributions. For example, what kind of underlying distribution would need to be modelled in order to output a course of action that would de-escalate military action, or would suggest reviewing the evidence for an attack formation?

As the Palantir platform extends into government courses of action in healthcare, logistics, border controls, immigration and asylum decisions, the political logic of underlying distributions promises to "leverage LLMs to enable a reasoning through each scenario and course of action" (Palantir, 2022b). In November 2023, Palantir were awarded the £350 million UK Government contract for an NHS "federated data platform". Though public concerns have understandably focused on access to personal health data by a US corporation, there ought to be deep concerns about what Palantir's generative models will do to actions and decisions in healthcare. It is precisely in this idea of "reasoning through" a possible course of action that the political logic of generative AI significantly shifts the form of political reason at work. The political logic of underlying distributions means that a course of action – whether military or clinical or bureaucratic – is immanent to a structure that is not intelligible as such. These distributions are no longer the relatively

[3] The relationship between a "target" and a "course of action" is complex and not reducible solely to the military target. Rather, the target in Palantir's model is more akin to what Samuel Weber calls "targets of opportunity" that coalesce the commercial logics of targeted advertising with the security logics of military targeting (2009; see also Aradau & Blanke, 2022, p. 1).l.

simple aggregates of political subjects encapsulated in the idea of populations, but rather link much more heterogeneous and granular ensembles of people and things. The politics of distributions in this generative sense differs from the important but now familiar criticism that models merely reflect or "parrot" the data on which they were trained (Bender et al., 2021). These models produce an ambiguous politics, in which the speculative – the probabilistic sampling of novel outputs – is generated and inferred from an assumed empirical – the heterogeneous data foundation on which these models are trained (Campolo & Schwerzmann, 2023). The political logic of the underlying distribution governs a world via the traversing of a data foundation so that decisions and courses of action will be immanent to the structure of the underlying distribution.

2. Latent space: the political logic of compression

Central to generative models' learning process is the capacity to compress input data into lower-dimensional representations, or a so-called *latent space*. The idea of dimensionality reduction significantly pre-dates generative AI in that it seeks to discard some features from large datasets to make possible "the classification, visualization, communication, and storage of high-dimensional data" (LeCun et al., 2015). However, the more recent idea of latency turns this reduction into a positive, productive force. Latent spaces first emerged as a crucial idea with the introduction of generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) in 2014 (Goodfellow et al., 2014). Broadly speaking, the models transform a data distribution into a compressed, lower-dimensional representation in order to learn salient patterns, attributes, and features within an input data distribution (Ruthotto & Haber, 2021, pp. 1–24). The implication of this hidden manifold is that, as some computer scientists have put it, "we now have access to an efficient, low-dimensional latent space in which high-frequency, imperceptible details are abstracted away", meaning that the model can now "focus on the important, semantic bits of the data" (Rombach et al., 2022, p. 4).

The etymology of latency, from the Latin, latentem meaning lying hidden, concealed or unknown, is suggestive of something that is hidden and unknown in a dataset, or crucially, in a broader population. It is precisely this capacity to surface to attention things that were otherwise hidden that animates latency as a political logic. As Kristen Veel (2021: 313) puts it, "the notion of *latency* is thus situated in the twilight zones between visibility and invisibility, knowing and unknowing, gesturing toward something hidden from view that we expect to emerge into visibility at some point in the future". The computational logic of a latent space – from which a model itself finds the most salient hidden features in data – begins to mutually resonate with a political logic that seeks out the latent hidden tendencies in populations, places, or scenes. It is a logic that promises that hidden tendencies in populations can not only be brought to the surface through an algorithmic compression of a data distribution but can also render populations tractable and governable in new ways.

While most commonly used in the context of algorithms generating images, the idea of the latent space nonetheless indicates the importance of algorithms learning through a process of *compressing* data. As Ilya Sutskever, co-founder and chief scientist at *OpenAI,* explains in an interview, "if you compress the data really well, you must extract all the hidden secrets which exist in it. Therefore, that is the key" (Sutskever, 2023). "Compression", he continues, "has the property that it discovers the secrets in the data. That's what we see with these GPT models [they] learn a compressed, abstract, useable representation of the world" (Sustkever, 2023; see also Deletang et al., 2023). Yet, this capacity of generative models to compress has also provoked recent debates regarding what precisely happens in the latent space and whether it is just producing a reductive and "blurry" representation of the world (see Chiang, 2023; Offert, 2021). From the perspective of these critics, the latent space demonstrates how generative models merely reduce and distort what is already known. However, if one takes seriously the logic of compression as a political rationality that mobilises the latent space, then it can never be simply a reduction or distortion of what the real world is. Rather, it is a highly generative, productive and derivative space of possibilities. In contrast to the reduction/distortion critiques, the latent space constitutes the world of the model: its limitations, its possibilities, its understanding, its constraints, its potentiality, its risks, and its promissory allure.

What kind of space is latent space? What is its political logic? Consider the example of a generative adversarial network (GAN). In an influential 2016 paper, machine learning researchers Alec Radford, Luke Metz, and Soumith Chintala sought to assess their Deep Convolutional GAN model and the quality of its output images. The researchers describe a method of evaluation called "walking in the latent space", explaining that:

> Walking on the manifold that is learnt can usually tell us about signs of memorization (if there are sharp transitions) and about the way in which the space is hierarchically collapsed. If walking in this latent space results in semantic changes to the image generations (such as objects being added and removed), we can reason that the model has learned relevant and interesting representations (Radford et al., 2016, p. 7).

Here, the latent space is not simply a question of reduction, distortion or loss of information. Rather, the model is "collapsing" the data in a highly organised and meaningful way. Walking the latent space becomes not only a way to explore what the model has learned but also to generate output samples that approximate the distribution of the training data. While it is constrained by the original training data (de Vries, 2020; Offert, 2021), the latent space indicates the extent to which the model learns latent factors about the world, learns what it considers to be important. It has developed and built a particular world.

The latent space also suggests that the model is capable of generating samples that variously fall between data points on which the model was trained. In the case of the facial image datasets used for biometric recognition algorithms, for instance, the generation of novel data points has significant political valance (Jacobsen, 2023). There are important critical interventions that point to the gendered and racialised "bias" that resides within the data on which facial recognition algorithms are trained (Buolamwini & Gebru, 2018). However, with the logic of latency it would be insufficient to address or correct the discriminatory datasets alone. Rather, latent space transforms the facial image datasets, generating synthetic faces that fall between data points and cannot be traced back to any specific data element. Thus, the racialised classification of faces becomes a matter of what Thao Phan and Scott Wark (2021) term "racial formations as data formations". These in-betweens and interstices problematise the ethico-political claim that one could retrieve or remedy the data points that are responsible for discrimination. A racialised or gendered output from a generative model may never be traceable since the interstitial samples do not strictly exist anywhere as data points in the training data: they are latent and learned.

The ethical and political stakes of latency stem from the twinned ideas we have described: the "hidden" secrets of compressed data and the generation of samples between data points. Where the technical logic of latent space foregrounds the hidden features harboured in a data distribution, its twinned political logic imagines a world where, given sufficient data and the power of compression, the hidden political solutions and resolutions can be found. Consider, for example, the UK Home Office's 2023 call for AI developers to participate in a three day "hackathon" in order to "search for ways to use AI to cut the asylum application backlog" (Gentleman, 2023). The government procurement of large language models for immigration and asylum decisions – implemented via a commercial hack event – explicitly referenced the desire to leverage the "huge existing database of thousands of hours of previous asylum interviews" to "identify trends". What is at stake is not merely the tracking forward or statistical inference of past cases to

future decision outputs. Rather, the large and language-based asylum datasets – once subject to compression and the discovery of latent features – construct in an unsupervised way the key factors at play in future potential decisions. The specific and singular ethico-politics of an asylum claim – even the keywords for violence, torture, or trauma captured within the text of past applications – is actively compressed in order to yield the latent or hidden tendencies.

Though it is not possible for us to observe directly what happens in something like an asylum or immigration latent space, it is clear from other experiments with LLMs that fundamental categories of race, nationality and ethnicity are reconfigured through the process of compression. For example, one group of Stanford researchers concluded that GPT3 had learned "latent concepts between examples in a prompt" (Xie et al., 2022, pp. 1–25, p. 1). When given the prompt "Albert Einstein was German \ n Mahatma Gandhi was Indian \ n Marie Curie was", the model inferred that Marie Curie was Polish. The model had not been explicitly pre-trained to know the nationality of Curie but had learned the latent concept of nationality between the examples. In many legal challenges to the use of protected characteristics by algorithms, the category of nationality has been successfully challenged and removed as a data input to government algorithms (Amoore, 2023). Significantly, the latent space *does not need* to include any specific data inputs for nationality, race, gender, or sexuality in order to "learn latent concepts between examples". More than political "proxies" that stand in place of something other, though, latency mobilises the space in-between plural data points (Mulvin, 2021). The political logic of the generative model values the unknown or latent features that "we cannot observe directly" and that are derived from the processes of compression (Goodfellow, Bengio & Courville, 2016: 67). If compression and dimensionality reduction are understood not simply to invoke "blurriness" in a negative sense but rather to imply a productive process that is suggestive of structures and concepts, then the latent space becomes also a political space for the governing of the latent tendencies of population. This is a model of the world that is not strictly an accurate or precise "fit" or picture of some actually existing set of relationships in the politics of asylum. Rather, the generative model is building a world of asylum and immigration (or warfare, biology of protein structures, or any domain) in which the relations are brought into being in and through the latent space.

This productive valuing of latent features also significantly problematises interventions that seek to govern generative models or to establish so-called "guardrails" to control their behaviour. The point is forcefully made in *OpenAI*'s own technical report on GPT-4. They write that although safety measures were put in place throughout its development, "the fundamental capabilities of the pre-trained model, such as the potential to generate harmful content, remain latent" (*OpenAI*, 2023, p. 68). In this stark assessment, the potential for *harm is latent* within the model precisely because latency disrupts any sense of a linear or causal relationship between input and output. Understood in this way, there are profound limits on the scope for modifying inputs to reduce the harms of generative AI. One could establish guardrails to limit toxic outputs, but yet the model retains the potential to generate something hidden and not otherwise present in the data – present only in the latent space. Features that would not otherwise be related or connected become associated in and through the latent space. Though these features are not linear-causal or input-output, they nonetheless lend renewed and novel force to the politics of profiling and "guilt by association" that seeks latent tendencies in populations, places, and scenes.

3. Sequences: the political logic of attention

The idea of the sequence as ordering device has characterised much of the history of algorithms, whether as a "recipe composed in programmable steps" or as a sequence of "if-then-else" rules, "subdivided into steps" so that "a machine could execute them" (Gillespie, 2016, p. 19; Bucher, 2018; Daston, 2022, p. 8). As advances in natural language processing began to be taken up by other "sequence oriented" tasks, such as human genome sequencing, the relationship between the sequence and the model began to transform. It is this transformation that interests us here. What has happened to the sequence as an ordering device with generative AI? What are the implications of new logics of the non-linear or broken sequence for the political logics of generative AI?

Sequences in NLP represent the order of words in an input sequence, so that a model can predict the next word in a sequence given a sequence of preceding words. Sequence modelling represents a means of *ordering* things, both in the *mathematical* sense of the arrangement of elements and in the *political* sense of ordered hierarchies or classifications (Devlin, 2000; Foucault, 2003, p. 289). Though generative AI has important origins in the linguistic and syntactical sequences of NLP, significantly it also breaks with the idea of the sequence as a linear left-to-right series of steps, expanding input sequences beyond immediate contexts, and parallelizing to allow attention to be paid to certain parts of sequences. The sequence is retained as a concept that orders a picture of the world and yet is exploded and broken into the parallel architecture of the transformer. It is this curious simultaneous retention and destruction of the sequence as ordering device that we observe to be shaping a broader political logic of speculative and predictive global dependencies.

In an interview in 2022, *OpenAI*'s Ilya Sutskever reflected on the breakthrough moments of successive GPT models. "The great discovery of GPT3", he explained, "is that predicting the next word in text is a very interesting objective" (Sutsekver, 2022). As he described it, the prediction of the next word in a sentence is a machine learning objective that yields much more than merely the most probable character or word in a natural language sequence. "If you can have a good enough guess at the next word in a text", he continued, "then it means you must understand the text". The promise of large language models is that predicting the next token in a sequence affords a capacity beyond the sequence itself: an understanding of the whole structure of the underlying text: "Next character prediction, next something prediction, has the special property that it discovers some hidden structure in the data".

As a set of propositions and assumptions about the world, these models embody a particular orientation to prediction and unknown future states. The futures-oriented act of predicting the next thing in a sequence – whether a DNA sequence, a protein structure, or a word in a sentence – necessarily involves knowledge that is greater than the sum of the parts of the sequence, that is understanding of the underlying data distribution. To better grasp the implications of LLMs and their prediction of something that comes next in a sequence, it is helpful to address how the sequence itself has functioned and mutated as an ordering device within the precursors to GPT, and how a specific political logic of attention has become powerfully installed in generative AI models.

Generative AI has significantly transformed and reconfigured the NLP concept of the linear sequence of tokens, breaking apart and parallelizing the relations in data. The genealogy of the sequence in contemporary AI models arguably begins in 2014 when significant experiments take place in terms of the *reversal or reordering of sequences*. While the 2012 AlexNet algorithm represented a major defining breakthrough in image-based machine learning (Krizhevsky et al., 2012), in 2014 a parallel breakthrough took place in language-based models, when deep neural networks began to be applied to unlabelled sequences. The so-called "sequence to sequence" (Seq2Seq) architecture was explicitly positioned as a break with the tightly drawn sequential logics of NLP and recurrent neural networks (RNNs) (Sutskever, Vinyals & Le, 2014). Existing approaches to NLP were described by the authors as having a "good sequence architecture" but one that was limited in its application "to problems whose inputs and targets can be encoded with vectors of fixed dimensionality", making it "difficult to reach back in the sequence" or to deal with "sequences whose lengths are not known a-priori" (Sutskever, Vinyals & Le, 2014: 2).

The political significance of these models is that the idea of the sequence begins to take on a new form that can deal with large volumes of long sequences, and to apply its knowledge in "a domain independent

way" (2014: 2). The capacity to deal with unencountered sequences across multiple domains (from text translation to DNA sequencing) thus begins to emerge in language models in the years following the AlexNet image models. Seq2Seq used the 'Long Short-Term Memory' (LSTM) architecture to process data sequentially whilst sustaining information across many timesteps in the sequence. Key to the claimed benchmark performance of the Seq2Seq model in English to French translation tasks is the experimental reversal of the order of words in the source sequence (English) and not in the target sequence (French). "The simple trick of reversing the words in the source sentence", the authors propose, is "the key technical contribution of the work", so that each word is spatially closer to its corresponding word in the target sentence. It is in the experimentation of the sequence-to-sequence model that we can begin to locate a first important step in the transformation of linear sequential orders in favour of the idea that reversals and folds in ordering yield something useful to the model's knowledge of the world.

If the novel sequence to sequence models began to map the possibilities for transformations in the order of the sequence, then the *attention* models that followed introduced the problem of *which parts* of the sequence are most relevant for a given task. The problem that inspired work on attention was the long sentences in machine translation and crowded images in caption generation, where the compression required to encode the large input resulted in the degradation of performance. The rise of the attention mechanism that has become so integral to today's LLMs was driven by a need to decide which units in a sequence mattered, which data was important, on what the model's attention should be focused. The attention mechanism signals a key moment in a long history of forms of attention in the world, where "the problem of attention becomes a fundamental issue" within the human sciences and the governing of populations (Crary, 2001, p. 13; see Pedersen, Albris & Seaver, 2021). From this perspective, the major breakthrough of transformer models (the "T" in GPT), with the paper "Attention is All you Need", reordered the regime of attentiveness within machine learning (Vaswani et al., 2017). The transformer model was said to have "radicalized" the use of attention in sequence-to-sequence language modelling, dispensing entirely with recurrence and convolution in favour of an ensemble of attention mechanisms. Thus, the transformer eschewed the sequential recurrence of earlier models, relying instead on what the authors term a "multi-headed, self-attention mechanism" in both the encoder and decoder (2017: 2). With self-attention, the linearity of the sequence is broken and, crucially, parallelized on GPUs (Rella, 2023). In Vaswani and colleagues' terms, the parallelization not only increases the speed and efficiency of the model but also "allows the model to jointly attend to information from different subspaces at different positions" (2017: 2).

The transformer architecture likewise has inescapable political implications as a means of dividing, partitioning, attending to, and acting upon the world (Rancière, 2007; Foucault, 1991). In dispensing with the relatively linear sequences of recurrence and convolution, transformers extend the computational logic of "attention is all you need" into a pervasive political logic that one only needs to have sufficient data and to know which parts of it are most important to attend to. In contrast with the recurrent models that precede them, transformers use parallel attention layers to attend to multiple things simultaneously (sequence to sequence, but also vector to sequence, and sequence to vector), structuring a model of the world where multiple potential non-linear causal links, or "global dependencies between input and output" can be drawn (Vaswani et al., 2017, p. 2). The attention mechanism allows for the modelling of global dependencies "without regard to their distance in the input or output sequences" (2017: 2) and so relationships will be surfaced from across a "global" or long-range picture of the underlying data structure. As a political logic, the global dependency structures attentiveness to relationships that might otherwise be overlooked.

Consider, for example, a genomics scientist explaining how transformer models are changing the shape of what is afforded their attention in genomics: "We wrote the rules for chess and the rules for Go, but we did not write the rules for biology, so in a sense we have moved to radical empiricism" (Genomics England, 2021). The transformer model he is using in the lab is precisely drawing global dependencies between the input and output on vast human genome datasets, so that the model is structuring attention to relationships that are not formulated in advance as axiom, rule, or hypothesis. The machine learning sense of "attention" being "all you need" begins to install itself as a broader political logic, marked by an active alignment between non-linear and parallelized computation, the combining of vast public and private datasets, and the non-linear and simultaneous problems that are apparently to be addressed by LLMs. In short, the transformation in the idea of the sequence – from the linear order of NLP to parallel "multi--head attentiveness" – actively shapes a political imaginary where non-linear "global" dependencies can be drawn in multiple spheres of life, from policing and criminal justice to oncology treatment pathways and military strategy. As opposed to a technocratic politics of rationalised specialisation, these models move between domains, redrawing political and epistemic boundaries as they move.

4. Pre-training/fine-tuning: towards a zero shot politics

If the political *order* of generative AI is one in which sequences are broken and *de*-linearized in favour of scattered, non-linear global dependencies, how might we describe the types of *objectives* and *tasks* addressed by these machine learning systems, and their orientation to action in the world? Such questions are complicated by the extraordinarily diverse range of generative models, covering almost any linguistic or symbolic activity, as well as sensory orders, most notably the production of images. Behind this apparent diversity, however, lies a logic that traverses these domains: that of pre-training (the source of the "P" in the GPT acronym) and fine-tuning.

We opened this essay with a discussion of LeCun's appeal to the "world model" as a flexible and adaptable model capable of acting upon any new entity not encountered in the training data. Contrary to LeCun's sense that generative AI does not have a model of the world, for us the political logics of generative models are precisely instituting a governing logic that actively builds a kind of transformer worldview. Central to this worldview is a computational logic of pre-training and fine-tuning that now aspires to what is called "zero shot" learning: that is, to be able to complete a task with no labelled data available for each new class. The advent of pre-training and fine-tuning in *computational practice* was a crucial condition of possibility for today's explosion of generative AI as simultaneously a *political practice* that seeks to act and decide on new problems in the absence of formal, labelled, "known" data: we call this a "zero shot politics". In order to explore this idea, we reflect here on how the pre-training and fine-tuning paradigm emerged, before turning to the zero shot political logic we suggest it instantiates in the world.

Pre-training a model involves selecting a simple, broadly applicable training objective, such as predicting the next element in a sequence, and training an unsupervised model using large amounts of unlabelled data. The weights from this pre-trained model are then used as a starting point to train a supervised, "fine-tuned" successor on a related, but more specialized task, often with a smaller amount of labelled data (Dai, Andrew, & Quoc, 2015). The tendency for a small number of very large, pre-trained language models to be deployed across domains has led some to refer to them as "foundation models," exhibiting a tendency to "homogenization." (Bommasani et al., 2022, p. 5). However, the political implications of such models are not quite captured by this tendency to homogenize differences and are closer to a longer historical movement towards adaptive models that are said to be "domain agnostic" (Ribes et al., 2019). To be domain agnostic – or more precisely to claim domain agnosticism – is to disavow the particularity of situated knowledge and to clear the ground for machine learning models in all aspects of life.

As is so often the case in machine learning, a theoretical understanding of pre-training and fine-tuning techniques emerged out of

engineering problems relating to learning from large unlabelled corpora (Dai, Andrew, & Quoc, 2015, 1). These problems were particularly acute in sequence learning and NLP, where large amounts of text data was available but problematic to leverage using supervised methods due to the difficulty of labelling it. Pre-training – deriving a set of initial weights based on large amounts of unlabelled data – proved to be an effective machine learning solution to many of these difficulties. In sum, the political promise of domain agnostic models was conjoined with the allure of leveraging large amounts of unlabelled data across domains of governing.

Architectural changes, notably transformers, prompted further refinement of pre-training – fine-tuning relationships. The first GPT model associated the pre-training phase with unsupervised generative modeling and the fine-tuning phase with supervised discriminative modelling, where relationships between data and label in light of a single task or objective are modelled in a more direct way. Transformer architectures enabled pre-training to capture much longer-range, linguistic dependencies, and they even began to outperform architectures that had been expressly designed for specific tasks (Radford et al., 2018, p. 2). Subsequent models, such as BERT, further expanded the types of possible dependencies that could be captured during pre-training, notably by using masking to incorporate bidirectional relationships (Devlin et al., 2019, p. 2). GPT-3 continued this trend by using a pre-trained model that dispensed with supervised fine-tuning entirely, relying on "scale"—both in terms of the 175 billion model parameters and the size of the training dataset (8 million documents)— to create remarkably robust pre-trained representations (Radford et al., 2018, p. 3). Instead of updating the weights of a fine-tuned model by using a new labelled dataset, these more general models are given a few instances of a desired behaviour, for instance, a sentence translated from one language into another (few-shot learning). Even more minimally, a model is given a natural language description of the desired task in the form of a *prompt*: so-called "zero-shot" learning (Brown et al., 2020, p. 3).[4]

Beyond the immediate observation that pre-training and fine-tuning have supplied flexible models that smooth the deployment of generative AI into multiple domains of social and political life, the political logic of pre-training and fine-tuning registers a deeper epistemic transformation. Zero shot learning opens onto a horizon of a generalised and domain agnostic *exploration* of potentials. The initial pre-training and fine-tuning paradigm altered the models' relationship between the universal and the particular, the general and the specific. Models such as BERT used a universal or general-purpose LLM trained on a large corpus, and then fine-tuned it on a new set of literature for the particular task domain. Hence, specialized language models for domains such as biology (BioBERT) and finance (FinBERT) became specialist language processing models that adjusted the general LLM for particular deployment in the world. More recent models propose to obviate the need for smaller domain-specific datasets usually used for fine-tuning, instead leveraging the language modelling objective itself to elicit desired behaviours on tasks demonstrated in natural language, in the form of *prompts*.

The idea of relinquishing a domain-specific dataset for fine-tuning has considerable political implications. For example, where an immigration algorithm would be derived from a pre-trained LLM but fine-tuned on data specific to the domain (with a concomitant critical focus on data bias and discrimination), a zero shot approach relinquishes situated data in favour of exploring potentials and prompting the model. The "few shot" and "zero shot" ambitions and desires of the transformer model begin to crystallise in the form of *prompting*: "instead of finetuning a separate language model checkpoint for each new task, one can simply "prompt" the model with a few input-output exemplars demonstrating the task" (Wei et al., 2023, p. 2; see also ****, 2023). Prompting, understood as feeding a set of natural language instructions to the LLM to guide the style and content of the output, is a form of nudging the model that influences the navigation of the algorithm in its data space. While transfer learning assumed moving between self-contained domain spaces, prompting is "*task location* in the model's existing space of learned tasks" (Reynolds & McDonell, 2021, p. 1, original emphasis). As a prompt engineer interviewed by the Washington Post argued, "you're exploring the multiverse of fictional possibilities, sculpting the space of those possibilities" (Harwell, 2023). Prompting is thus a form of experimentation and exploration because it actively directs the behaviour of the LLMs towards portions of its own training dataset and away from others, but also towards portions of the "world" of data and away from others.

The rise of the logic of "zero shot" and exploratory task specification as prompting has profound implications. These logics are finding their way into political decisions, for example in the trialling of an LLM "AI red box" by the UK government's Cabinet Office, where generative AI will write draft responses to parliamentary questions (Fisher, 2024). The politics of zero shot, however, is not limited to the deployment of LLMs in political decisions, but rather it reconfigures the space of the political and the navigation of that space. When a generative model is used in government, the political decision of whether to close a hospital or to address child poverty, for example, will take place in a space that is always already shaped and delimited by the algorithm. Again, it is not only that language models "merely" predict the next word in a sequence – a sort of reductionism – but rather that the universalizing scale of zero-shot enables new ways of formulating tasks linguistically. Unlike the explicit command or the rigidity of an ethical rule, prompts work by probabilistically *eliciting* desired behaviours from models in light of a user's objective. New problems, like hallucinations, emerge, which look very different from the reflection of *biased* inputs. This well-known tendency of generative models to produce erroneous responses *not* included in their training set is in fact determined by their pre-training objective, which demands only the most probable response to an input sequence or prompt (Ouyang et al., 2022). Unlike supervised models, which are meant to reflect the "ground-truth" of a data-label relationship (Jaton, 2017), pre-trained language models may not be constrained by model-world correspondences in a factual sense. Here, the relationship between truth and ethics becomes blurred, as outputs are determined in a more self-referential way, through the estimation of distributions using large amounts of unlabelled data, rather than some correspondence between data and label or world.

With zero-shot learning, then, a logic of data and labels, and even fine-tuning, gives way to the use of linguistic prompts to model and shape tasks and actions in the world, a means of governing and responding to unknown or unencountered situations precisely on the basis of the underlying distributions, latent spaces, and broken linguistic sequences we have described. A zero shot politics harnesses and mobilises the technical force of generative AI – the unseen task; the corpus of unlabelled data; the domain agnostic model; the prompt that elicits desired behaviour – and forges a governing logic for a world of the unseen, the unknown and the unencountered.

5. Conclusions: a politics of world models

There can be little doubt that the advent of generative models is having profound consequences for some of the most fundamental political dimensions of our world. Indeed, the public debates have variously witnessed calls for a "pause" or a "moratorium" on the development of generative models (Paul, 2023). Notwithstanding the immediate controversy of the lead developers of generative AI displacing its present political responsibility into projections of future harms,

[4] The emergence of the 'prompt' in generative AI is itself a significant and complex genealogy (Burkhardt & RiederB, 2024). Though it is beyond the scope of this paper, the prompt is significant because it erodes the distinction between training a model and using a model – the "prompt" simultaneously guiding and refining the model and providing the means of using or interacting with the model.

such responses are manifestly insufficient; they sustain a separation between the model and the world, where the generative model remains an external object to be controlled or prohibited via the ban or the pause.

In this essay, beginning from a claim that generative models are already instantiating a model of the world, we elaborate a set of political logics and governing rationalities that thoroughly entangle and enmesh the computational model with a durable political model. From our perspective, one could imagine a world where, even with serious and defined constraints on generative AI as technology, the potentially plural harms of generativity as political logic will continue to proliferate. Of course, there are certain risks involved in taking seriously the world-making capacities of generative models, not least the risk that we engage in precisely the kind of hype that characterises the commercial and computational claims for LLMs. However, we need to excavate the technical architectures of generative AI precisely because they give rise to concepts – such as distribution, latency, prediction, zero-shot – that also build a political architecture that may outlive the technical model itself. Our aim, then, is to expand the landscape of critique, beyond the focus on inputs and outputs and into the political arrangements and propositions of the model itself.

In tracing a broader genealogy of four key aspects of generative AI, we have focused on why some concepts and techniques have become more enduring forms of knowledge that resonate across *both* computational *and* political logics. With *generativity*, a particular form of knowledge of the world – drawing upon historical statistical antecedents of probability distributions for governing populations (Joque, 2022) – mobilises the "underlying distribution" to generate "courses of action" from multiple data sources, ranging from images of things to language itself. The LLM-derived courses of action are better understood to be contingent probabilities that nonetheless become political decisions that radically foreclose alternatives. With *latency*, the computational latent space – where a model finds the salient hidden features via the compression of data – becomes a seductive political promise of locating the hidden tendencies in populations, places or scenes. In contrast to the idea that compression produces degraded copies of training data, we highlight the logic of compression conceived as a productive and generative process. With new machine learning approaches to *sequences*, the capacity to predict the next thing, the next anything, becomes a broader logic of ordering the world. The transformer breakthrough with "attention is all you need" parallelizes these sequences so that non-linear or distant dependencies can be made to matter. It is a speculative and predictive political order that promises to govern non-linear problems by attending to the "global dependencies" in all available unlabelled data. Finally, with the *pre-training* and *fine-tuning* paradigm, the desire for "zero shot" learning (classifying the previously unencountered) becomes simultaneously desired as a zero-shot politics of responding to all unencountered events. With the rise of the prompt, the generative model offers itself as an open-ended and experimental mode of governing, with each prompt seeking to elicit desired behaviours from model and from the world.

To take seriously the already existing and emerging political logics of generative AI is not the same as assuming a congruence between the computational and the political logics at work. Indeed, a significant political space for intervention dwells precisely in the gaps, tensions and slippages between and within the computational and the political logics. There are gaps, many of them, and among them the gaps between the power arrangements of the political logic and the actual material capacities of the technologies of generative AI. For example, the very many failings and shortcomings of transformers – the problems of "out of domain" data, the errors in compositional tasks – signal the potentials that are not actualised, or what elsewhere has been called "the unattributable" that evades foreclosure (Amoore, 2020). In this way, the epistemic forces of generative AI point to a different horizon of power and resistance: where the claim to an *underlying distribution* must contain the possibility of *multiple other distributions*, or points on the distribution that were discarded alternatives.

CRediT authorship contribution statement

Louise Amoore: Writing – original draft. **Alexander Campolo:** Writing – original draft. **Benjamin Jacobsen:** Writing – original draft. **Ludovico Rella:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research has received funding from the European Research Council (ERC) under Horizon 2020, Advanced Investigator Grant ERC-2019-ADG-883107-ALGOSOC.

References

Amoore, L. (2020). *Cloud ethics: Algorithms and the Attributes of Ourselves and others, duke*. Duke University Press.
Amoore, L. (2023). Machine learning political orders. *Review of International Studies, 49* (1), 20–36.
Aradau, C., & Blanke, T. (2022). *Algorithmic reason: The new government of self and other*. Oxford: Oxford University Press.
Barry, A. (2001). *Political machines: Governing a technological society*. London: Athlone.
Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models Be too big?. In *In Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922, 610–23. FAccT '21.
Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homeogenization?. available at: https://arxiv.org/pdf/2211.13972.pdf.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Neelakantan, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
Bucher, T. (2018). *If…Then: Algorithmic Power and politics*. Oxford: OUP.
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15.
Burkhardt, S., Rieder, B., & B. (2024). Foundation models are platform models: Prompting and the political economy of AI. *Big Data & Society, 11*(2). https://doi.org/10.1177/20539517241247839
Campolo, A., & Schwerzmann, K. (2023). From rules to examples: Machine learning's type of authority. *Big Data & Society, 10*(2). https://doi.org/10.1177/20539517231188725
Chiang, T. (2023). ChatGPT is a blurry JPEG of the web. *The New Yorker*. Available from: https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.
Crary, J. (2001). *Suspensions of perception: Attention, spectacle, and modern culture*. MIT Press.
Dai, Andrew M., & Le, Quoc V. (2015). *Semi-supervised sequence learning*. ArXiv https://arxiv.org/abs/1511.01432.
Daston, L. (2022). *Rules: A short history of what we live by*. Princeton NJ: Princeton University Press.
DeLanda, M. (2011). *Philosophy and simulation: The emergence of synthetic reason, New York: continuum*.
Deletang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., et al. (2023). Language modelling is compression. *ArXiv*, 1–16.
Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society, 8*. https://doi.org/10.1177/20539517211035955. no. 2.
Devlin, K. (2000). *The math gene, London: weidenfeld and nicholson*.
Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. available at: https://arxiv.org/abs/1810.04805.
deVries, K. (2020). You never fake alone: Creative AI in action. *Information, Communication & Society, 23*(14), 2110–2127.
Fisher, L. (2024). *UK Government to trial "red box" AI tools*. Financial Times, 28 February 2024.
Foucault, M. (1991). Governmentality. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect: Studies in governmentality*. Chicago IL: Chicago University Press.
Foucault, M. (2003). *The order of things*. London: Routledge.
Foucault. (2007). *Security, Territory, Population: Lectures at the Collège de France 1977-78*. Basingstoke: Palgrave Macmillan.
Gentleman, A. (2023). UK government "hackathon" to search for ways to use AI to cut asylum backlog. *The Guardian 29 April*.

Genomics England. (2021). Genomics in AI/ML. In *paper presented at the NVIDIA GTC conference, November 8-11, 2021*. Video available at: https://www.youtube.com/watch?v=r_mh2krxDoc. Last accessed May 2024.
Gillespie, T. (2016). Algorithm. In B. Peters (Ed.), *Digital keywords: A vocabulary of information society and culture* (pp. 18–30). Princeton NJ: Princeton University Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.
Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *arXiv*. https://doi.org/10.48550/arXiv.1406.2661
Gordon, C. (1991). Govermental Rationality: An Introduction. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault Effect: Studies in Governmentality* (pp. 1–52). Chicago: Chicago University Press.
Harwell, D. (2023). 'Tech's hottest new job: AI whisperer. *No Coding Required*. *Washington Post* 25 February 2023 https://www.washingtonpost.com/technology/2023/02/25/prompt-engineers-techs-next-big-job/.
Hayles, N. K. (2023). Subversion of the human aura: A crisis in representation. *American Literature, 95*(2), 255–279.
Jacobsen, B. N. (2023). Machine learning and the politics of synthetic data. *Big Data & Society, 10*(1).
Jaton, F. (2017). We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science, 47*(6), 811–840. https://doi.org/10.1177/0306312717730428.
Joque, J. (2022). *Revolutionary mathematics: Artificial intelligence, statistics and the logic of capitalism*. London: Verso.
Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 2*, 1097–1105.
Law, J. (1993). *Organizing modernity: Social ordering and social theory*. Oxford: Wiley-Blackwell.
LeCun, Y. (2022). A path towards autonomous machine intelligence. Open Review. Available at: https://openreview.net/pdf?id=BZ5a1r-kVsf Last accessed May 2024.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.
McQuillan, D. (2018). Data science as machinic neoplatonism. *Philosophy and Technology, 31*(2), 253–272.
Mol, A. (2003). *The Body Multiple: Ontology in Medical Practice*. Durham NC: Duke University Press.
Morgan, M. (2012). *The world in the model: How economists work and think*. Cambridge: Cambridge University Press.
Mulvin, D. (2021). *Proxies: The cultural work of standing in*. London: Penguin.
Ng, A., & Jordan, M. (2001). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (Vol. 14)MIT Press.
Offert, F. (2021). Latent deep space: Generative adversarial networks (GANs) in the sciences. *Media and Environment, 3*, 2. https://doi.org/10.1525/001c.29905
OpenAI. (2023). *GPT-4 technical report. ArXiv*, 1-98.
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., et al. (2022). Training Language models to follow instructions with human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2203.02155
Palantir. (2022a). Artificial intelligence platform (AIP) defense and military. *Demo*. available at: https://youtu.be/XEM5qz_HOU. (Accessed 6 June 2023).
Palantir. (2022b). Introducing Palantir AIP: Capabilities and product demo. https://www.youtube.com/watch?v=Xt_RLNx1eBM. (Accessed 23 October 2023).
Paul, K. (2023). Letter signed by Elon Musk demanding AI research pause sparks controversy. The Guardian. Available at: https://www.theguardian.com/technology/2023/mar/31/ai-research-pause-elon-musk-chatgpt Accessed(1 April 2023).
Pedersen, M. A., Albris, K., & Seaver, N. (2021). The political economy of attention. *Annual Review of Anthropology, 50*(1), 309–325.
Phan, T., & Wark, S. (2021). Racial formations as data formations. *Big Data and Society, 8*, 2. https://doi.org/10.1177/20539517211046377
Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv*, 1–16.
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language understanding by generative pre-training. *Open*. available at: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. October 2023.
Rancière, J. (2010). *Dissensus: On politics and aesthetics, New York: continuum*.
Rancière, Jacques (2007). *The Future of the Image*. London: Verso.
Rawls, J. (1971). *A theory of justice*. New York: Belknap Press.
Rella, L. (2023). Close to the metal: Towards a material political economy of the epistemology of computation. *Social Studies of Science*. https://doi.org/10.1177/03063127231185095. Online first.
Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv*. http://arxiv.org/abs/2102.07350.
Ribes, D., Hoffman, A. S., Slota, S. C., & Bowker, G. (2019). The logic of domains. *Social Studies of Science, 49*(3), 281–309.
Rombach, R., Blattmann, A., Lorenz, A. D., Esser, P., Ommer, B., & B. (2022). High-resolution image synthesis with latent diffusion models. *ArXiv*, 1–45.
Rubinstein, Y. D., & Hastie, T. (1997). Discriminative vs informative learning. In *Proceedings of the third international conference on knowledge discovery and data mining* (pp. 49–53). Newport Beach, CA: AAAI Press. KDD'97.
Ruthotto, L., & Haber, E. (2021). *An introduction to deep generative modeling' GAMM – mitteilungen*.
Seaver, N. (2018). What should an anthropology of algorithms do? *Cultural Anthropology, 33*(3), 375–385.
Sutsekver, I. (2022). Foundation models. In *Lecture delivered at Stanford HAI spring conference, 12 April 2022*. Available at: https://www.youtube.com/watch?v=W-F7chPE9nU. June 2023.
Sutskever, I. (2023). AI today and the vision of the future. In *Interview by jensen huang, nvidia GTC conference, 27 march 2023*. https://www.youtube.com/watch?v=-yquJiNKlAE. May 2023.
Sutskever, I., Vinyals, O., & Le, Q. (2014). *Sequence to sequence learning with neural networks*.
Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., ones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need'*.
Veel, K. (2021). Latency. In N. Bonde Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), *Uncertain archives: Critical keywords for big data* (pp. 313–321). Cambridge MA: MIT Press.
Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality*. New York: Basic Books.
Weber, S. (2009). *Targets of opportunity*. New York: Fordham.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought prompting elicits reasoning in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2201.11903
Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). *An explanation of in-context learning as implicit bayesian inference. ArXiv*.